A MIXED SINUSOIDALLY EXCITED LINEAR PREDICTION CODER AT 4 KB/S AND BELOW

Suat Yeldener[*]*, Juan Carlos De Martin, and Vishu Viswanathan*

DSP Solutions R & D Center, Texas Instruments, Dallas, Texas, USA
E-mail: [demartin|vishu]@csc.ti.com

ABSTRACT

There is currently a great deal of interest in the development of speech coding algorithms capable of delivering toll quality at 4 kb/s and below. For synthesizing high quality speech, accurate representation of the voiced portions of speech is essential. For bit rates of 4 kb/s and below, conventional Code Excited Linear Prediction (CELP) may likely not provide the appropriate degree of periodicity. It has been shown that good quality low bit rate speech coding can be obtained by frequency domain techniques such as Sinusoidal Transform Coding (STC) [1], Multi Band Excitation (MBE) [2], Mixed Excitation Linear Prediction (MELP) [3], and Multi-Band LPC (MB-LPC) [4] vocoders. In this paper, a speech coding algorithm based on an improved version of MB-LPC is presented. Main features of this algorithm include a multi-stage time/frequency pitch estimation and an improved mixed voicing representation. An efficient quantization scheme for the spectral amplitudes of the excitation, called Formant Weighted Vector Quantization, is also used. This improved coder, called Mixed Sinusoidally Excited Linear Prediction (MSELP), yields an unquantized model with speech quality better than the 32 kb/s ADPCM quality. Initial efforts towards a fully quantized 4 kb/s coder, although not yet successful in achieving the toll quality goal, have produced good output speech quality.

1. INTRODUCTION

The traditional speech coding systems are based on a simple underlying model that uses a binary, voiced/unvoiced excitation source and a time varying synthesis filter. Although vocoders of this type are capable of producing intelligible speech output, they have not been successful in synthesizing natural, high quality speech. In addition, these vocoders have provided a performance which degrades rapidly in the presence of background noise and which saturates at higher bit rates. There has been a lot of research devoted recently to improving these types of vocoders. The improvements have been based primarily on better modeling and quantization of the excitation signal after removal of short and long term correlations. However, these improvements result an increased bit rate for high quality speech output. Both CELP and MBE vocoders are capable of producing good quality speech at around 4.8 kb/s. Below 4.8 kb/s, however, these coders suffer from the distortion introduced by coarse quantization of model parameters due to the limited number of bits.

An alternative approach is to apply the sinusoidal model to the excitation signal in a linear prediction-based system. This approach has already shown the potential of producing good quality speech at low bit rates (4 kb/s and lower). It exploits the advantages of both time domain and frequency domain techniques to improve the speech quality at very low bit rates. The new U.S. Federal Standard speech coder operating at 2.4 kb/s [5] can be viewed as an example of this approach. It was preferred over several other speech coding algorithms including the Advanced Multi Band Excitation (AMBE), the Prototype Waveform Interpolation (PWI), and the Sinusoidal Transform Coding (STC). Also, the 2.7 kb/s MB-LPC coder reported in [6] produced consistently better speech quality than the 4.15 kb/s Inmarsat-M IMBE speech coder. In this paper, an improved version of the MB-LPC algorithm, operating at 4 kb/s and below, is described, and performance results are provided.

2. MSELP SPEECH MODEL

The simplified block diagram of our sinusoidally excited speech coding algorithm is shown in Fig. 1. In this algorithm, we use a speech production model where speech is formed as the result of passing an excitation, $e(n)$, through a linear time-varying LPC filter that models the spectral shape of the speech spectrum. At the encoder, a frame of speech is first analyzed to obtain speech production or LPC filter parameters. This filter is represented by 10 LPC coefficients, which are quantized in the form of Line Spectral Frequency (LSF) parameters. The quantized LPC filter coefficients are then used for LPC inverse filtering of the speech signal to generate the residual signal, which has a relatively flat spectrum. The excitation model parameters are then obtained, using the original speech to estimate the pitch and the mixed voicing information and using the residual signal to compute the spectral amplitudes.

Accurate pitch estimation has remained the most difficult problem in speech signal processing for decades. Some pitch estimation algorithms were found to produce good performance for some input conditions and some others for different input conditions, but it is very difficult to find one that produces consistently good results for a variety of input speech conditions. Therefore, we are proposing a multi-stage pitch estimation algorithm that is robust and provides accurate pitch for a variety of input conditions. In this algorithm, the whole pitch range is divided into various subranges and an optimal pitch candidate for each sub-range is chosen. The method to choose an optimal pitch candidate is to use a simple pitch cost function. Since the final pitch will not be computed directly at this stage, which cost function is used is not very important; therefore, we chose a computationally efficient one to obtain pitch candidates. In our experiments, the frequency domain approach reported in [7] is used. The next step in pitch estimation procedure is to compute an average pitch value using previous

[*] Suat Yeldener is now with COMSAT Laboratories, Clarksburg, MD, USA. Email: yeldener@ctd.comsat.com.

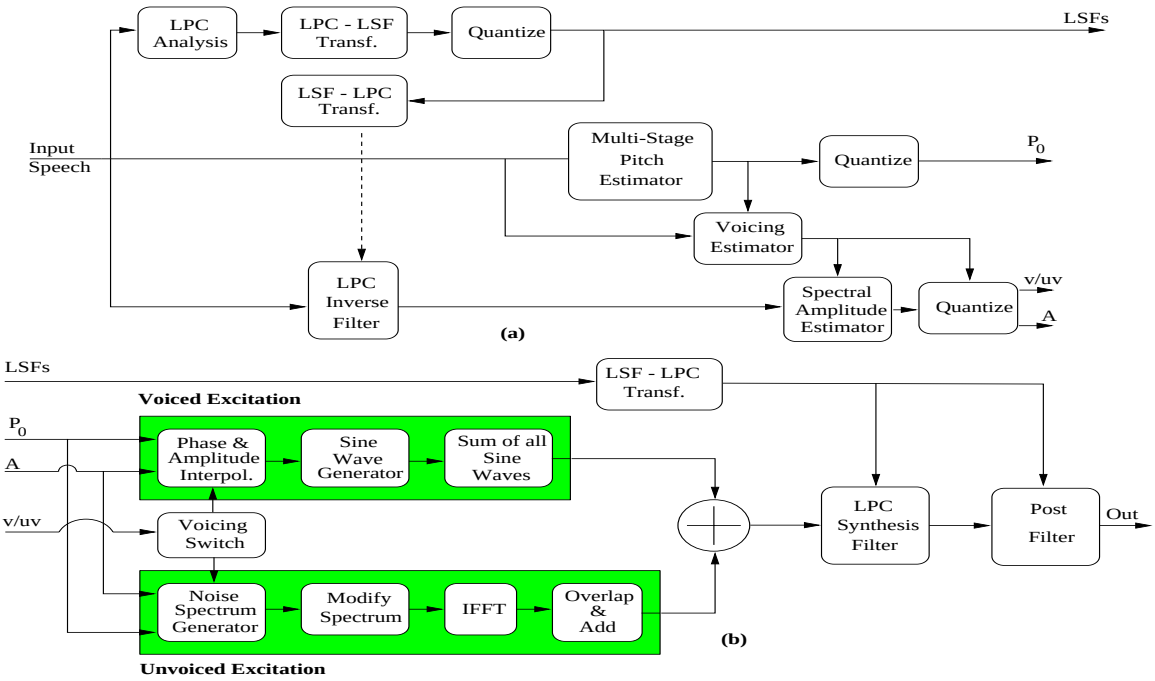

Figure 1: Simplified block diagram of speech coder: (a) encoder; (b) decoder.

pitch periods. The average pitch is used to switch between two pitch estimation algorithms: time and frequency domain analysis by synthesis pitch estimation. The idea here is that, during short pitch periods, there are just a few harmonics and it would be easier to match the time domain waveforms than their spectra; for long pitch periods, it would be the other way around. Therefore, an average pitch period is used to decide which algorithm (time domain or frequency domain) will be used to estimate the final pitch period. The block diagram of the time-domain analysis-by-synthesis pitch estimation algorithm is shown in Fig. 2.

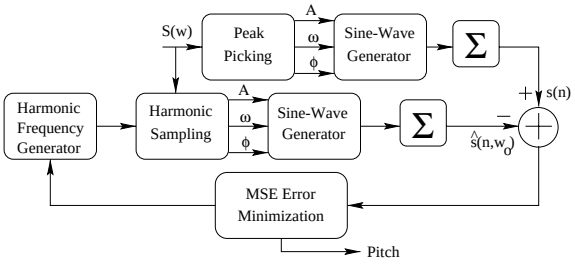

Figure 2: Time domain ABS pitch estimation algorithm

In this algorithm, a peak picking function is applied to obtain the peak spectral magnitudes and then the sine waves corresponding to these peaks are generated and added together to form the reference speech signal. For each fundamental frequency candidate, the speech spectrum is sampled at the harmonic frequencies and the sine waves for each harmonic are generated and added together to form the synthesized speech signal. The two signals are then compared, and the Mean-Squared Error (MSE) is minimized by choosing the best pitch candidate.

The frequency-domain analysis-by-synthesis pitch estimation algorithm follows a similar approach, but in the frequency domain. The block diagram of the frequency domain ABS pitch estimation algorithm is shown in Fig. 3.

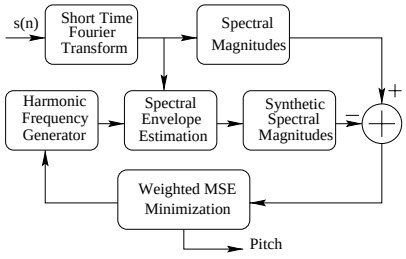

Figure 3: Frequency domain ABS pitch estimation algorithm

A short-time Fourier transform of each speech segment is computed and the resulting spectral magnitudes are kept as reference. For each fundamental frequency candidate, a spectral envelope is computed using the original speech spectrum. This envelope is then used to reconstruct the synthetic spectral magnitudes, which are compared to the reference. In this case, since the low-frequency components are more important than the high-frequency ones, a weighted MSE is computed and then minimized for each candidate to obtain the optimal pitch.

After estimating the pitch, the spectral amplitudes of the residual are then obtained by searching the spectral peaks around each harmonic lobe.

Voicing information is another factor that influences the per-

formance of a speech model. In the traditional speech analysis-synthesis systems, the excitation source generally uses a pitch period and a binary voiced/unvoiced decision for the entire speech frame. In Multi Band Excitation (MBE) vocoder [2], a different approach is taken to represent the voicing information. It is assumed that the entire speech frame is composed of both voiced and unvoiced excitation components. As a result of this, the speech spectrum is divided into various frequency bands and a binary decision for each band is made. This improves the modeling of the excitation signal over conventional vocoding techniques. However, whenever there is a voicing error, especially in low frequency bands, this results in perceptually objectionable degradation in the quality of the output speech. Therefore, recently we reported in [6] that the voicing information can be represented by a cut-off frequency that separates the voiced (low frequency components) and the unvoiced (high frequency components) portions of the speech spectrum. Representing voicing information this way is an efficient way to represent mixed type of speech signals.

As further improvement, in our proposed approach we allow all the harmonics of the fundamental to be composed of both voiced and unvoiced energy. A voicing probability as a function of frequency, $P_v(f)$, is used to define the ratio between voiced and unvoiced harmonic energies. This voicing model requires a large number of bits to represent the voicing probabilities for all harmonics. To resolve this problem, we have developed a simplified model using the idea of cut-off frequency and a constant voicing function for higher frequency harmonics. As a result, low frequency components up to the cut off are purely voiced and the harmonics above the cut-off are mixed having both voiced and unvoiced energies for each harmonic. The typical voiced and unvoiced probability curves are shown in Fig. 4.

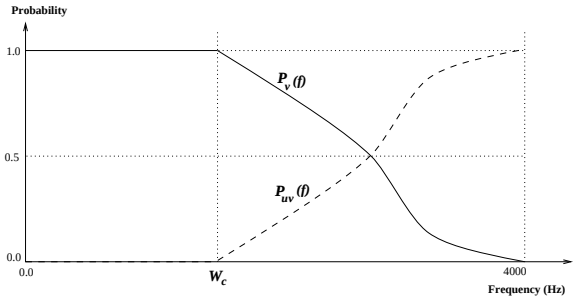

Figure 4: Typical voiced and unvoiced probability functions

At the decoder, the voiced part of the excitation waveform is determined as the sum of harmonic sine waves. For the unvoiced part of the excitation spectrum, a white random noise spectrum modified using the decoded spectral amplitudes is used for the unvoiced frequency region. The voiced and unvoiced excitation signals are then added together to form the overall synthesized excitation signal. The resultant excitation is then shaped by the LPC filter to form the synthesized speech. To enhance the output speech quality, a frequency domain post-filter is used [4].

3. INITIAL 4 KB/S CODER DESIGN

Straightforward quantization and coding of each of the MSELP model parameters would result in a good coder; however, the transmission rate would be too high. Therefore, efficient schemes to quantize the model parameters are necessary to achieve a good 4 kb/s speech coding algorithm. A frame size of 10 ms would be a good choice at 4 kb/s speech coding as it is compatible with ITU-T 4 kb/s speech coding requirements. In this case, 40 bits per 10 ms frames are available for 4 kb/s rate. In our design, we grouped two 10 ms frames together and transmitted only once to gain efficiency in quantizing the model parameters at the desired bit rate. The bit allocation among the model parameters for an overall bit rate of 4 kb/s with 10 ms frame length is given in Table 1.

Parameters	*Bits for* $(m-1)^{th}$ *Frame*	*Bits for* $(m)^{th}$ *Frame*	*Bit Rate (kb/s)*
Pitch	8	5	0.65
LSF Coef.	3	29	1.60
Exc. Gain	10	0	0.50
Spec. Shape	11	11	1.10
Voicing	0	3	0.15
Total	32	48	4.0

Table 1: Bit allocation for 4 kb/s coder.

The pitch of each speech frame is searched from 16 to 128 samples and quantized using 8 bits with half-sample accuracy for every other speech frame. The adjacent frame pitch values are then quantized differentially using 5 bits. The 10 LSF coefficients for $(m-1)^{th}$ frame are quantized using 5-stage vector quantization with 29 bits. The bit allocation of individual stages is $\{6, 6, 6, 6, 5\}$. The LSF coefficients for the m^{th} frame are quantized using the concept of optimal linear interpolation. In order to obtain the best performance, an attempt is made to minimize the Mean Squared Error (MSE),

$$E_k = \sum_{i=0}^{p} [l_m(i) - lsf_k(i)]^2 \quad (1)$$

where p is the LPC order, $l_m(i)$ are the original LSFs for the m^{th} frame and

$$lsf_k(i) = l_{m-1}(i) + [l_{m+1}(i) - l_{m-1}(i)] \frac{k}{M-1} \quad ; 0 \le k < M. \quad (2)$$

are the interpolated LSF's, m denotes the current frame index, and M is an integer that is a power of 2. The M set of interpolated LSF coefficients are then compared with the original LSF coefficients. The index for the best interpolated LSF coefficients, $k_{best} = k$, which minimizes the MSE, E_k, is then coded and transmitted using 3 bits. The voicing cut-off frequency is quantized using 3 bits for $(m-1)^{th}$ frame and is not transmitted for the m^{th} frame; the latter value is obtained by linearly interpolating the adjacent cut-off frequencies. The excitation gains for $(m-1)^{th}$ and m^{th} frames are grouped together and vector-quantized using a 10-bit codebook. The spectral amplitudes are quantized using a method called Formant Weighted Vector Quantization. Since the number of harmonics varies from one frame to another, the harmonic spectral amplitudes are linearly interpolated to form a fixed vector dimension. Since the low-frequency harmonics are perceptually more important than the high-frequency harmonics, the MSE is weighted by giving more emphasis to low-frequency components

during codebook training. During quantization of the spectral amplitudes, each codevector is down-sampled by the pitch and then a formant weighted VQ,

$$w_f(k\omega_0) \;=\; w(k\omega_0)\left(\frac{H(k\omega_0)}{F(k\omega_0)}\right)^{\gamma} \qquad (3)$$

which gives more emphasis to formant amplitudes, is applied in the computation of MSE and hence in the selection of the best codevector. In Eq. 3, $H(k\omega_0)$ and $F(k\omega_0)$ are the frequency responses of the LPC synthesis filter and the linearly interpolated formant peaks, respectively, sampled at the harmonics of the fundamental frequency and

$$w(\omega) \;=\; \left[1.0 \;-\; \left(\frac{\beta\omega}{N}\right)\right]^{\alpha} \quad ; \quad 0 \le \omega < N \qquad (4)$$

is the constant weighting function that gives more emphasis to low frequency components. In Eq. 4, α and β are fractional constants. In our experiments, we used $\alpha = 0.8$ and $\beta = 0.25$. A 2-stage vector quantization is applied to the spectral shape using 11 bits ($\{6, 5\}$ bits).

For bit rates of below 4 kb/s, the frame size may be increased, keeping the quantization schemes same as above. In this case, there is not much degradation in speech quality when a frame size of 15 ms is used, which leads to a bit rate of 2667 b/s.

4. SPEECH QUALITY TESTING

To verify the performance of the MSELP model, we ran an informal pair-wise listening test comparing the unquantized mixed sinusoidally excited LP model and the 32 kb/s ADPCM. For this test, each sentence was processed by our speech model and by the 32 kb/s ADPCM and the sentence pairs were presented to the listeners in a randomized order. The listeners rated the speech quality using an absolute scale ranging from -2 to 2, as shown in Table 2.

Score	*Description*
2	Coder A is better than Coder B
1	Coder A is slightly better than Coder B
0	Coder A is similar to Coder B
-1	Coder A is slightly worse than Coder B
-2	Coder A is worse than Coder B

Table 2: The rating scale used in the listening test

In the test we used speech from 3 female and 4 male speakers, for a total of 36 sentence pairs. Six listeners performed the test. The overall test results are shown in Table 3.

The data in Table 3 shows that, whenever a preference was expressed, 80% of the time it went to MSELP, suggesting that it performed better than the 32 kb/s ADPCM. This shows that the mixed sinusoidally excited speech model could provide high quality speech.

Also, 4 kb/s MSELP and 32 kb/s ADPCM coders were tested at a variety of input conditions using the ACR and CCR experiments defined by ITU-T for 4 kb/s standardization. However, the 4 kb/s MSELP coder did not meet the requirements of ITU-T. Further work is warranted to improve the speech quality performance of the MSELP coder at 4 kb/s.

	Preferences		
Score	**No of Votes**	**%**	*Pref. Coder*
2	35	16.2	MSELP (Strong)
1	86	39.8	MSELP
0	65	30.1	No Pref.
-1	26	12.1	ADPCM
-2	4	1.8	ADPCM (strong)

Table 3: Paired comparison test results between the unquantized MSELP speech coder and the 32 kb/s ADPCM coder.

5. CONCLUSION

In this paper, a mixed sinusoidally excited LP coding algorithm operating at 4 kb/s and below was presented. Methods for a multi-stage pitch estimation and a mixed voicing representation were also described. For quantization of excitation spectral amplitudes, a formant-weighted vector quantization technique was used, providing efficient quantization. An informal subjective listening test was conducted and the results indicate that the unquantized MSELP speech model produces better speech quality than 32 kb/s ADPCM under clean input speech condition. Initial efforts towards a fully quantized 4 kb/s coder, although not yet successful in achieving the toll quality goal, have produced good output speech quality.

6. REFERENCES

[1] R. J. McAulay, T. F. Quatieri, "Speech Analysis/Synthesis Based on Sinusoidal Representation," IEEE Trans. ASSP, 1986, Vol. 34, p. 744–754.

[2] D. W. Griffin, J. S. Lim, "Multi Band Excitation Vocoder," IEEE Trans. ASSP, 1988, Vol. 36, No. 8, p. 664–678.

[3] A. V. McCree, T. P. Barnwell, "A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding," IEEE Trans. Speech & Audio Processing, 1995, p. 242–250.

[4] S. Yeldener, A. M. Kondoz, B. G. Evans, "Multi-Band Linear Predictive Speech Coding at very Low Bit Rates," IEE Proc. Vis. Image and Signal Processing, Vol. 141, No. 5, October 1994, p. 289–296.

[5] A.V. McCree et al., "A 2.4 Kb/s MELP Coder Candidate for The New U.S. Federal Standard," Proc. ICASSP, 1996, p. 200–203.

[6] S. Yeldener, A.M. Kondoz, B.G. Evans, "A High Quality Speech Coding Algorithm Suitable for Future Inmarsat Systems," Proc. 7th European Signal Processing Conf. (EUSIPCO-94), Edinburgh, September 1994, p. 407–410.

[7] R. J. McAulay and T. F. Quatieri, "Pitch Estimation and Voicing Detection Based on a Sinusoidal Speech Model," Proc. ICASSP, 1990, p. 249–252.